РЕАЛИЗАЦИЯ СИНТАКСИЧЕСКОГО ГЕНЕРАТОРА ПРЕДЛОЖЕНИЙ ТАТАРСКОГО ЯЗЫКА

А.Р.Гатиатуллин

Казанский государственный университет
Djavdet.Suleymanov@ksu.ru

Введение. В программный комплекс АРМ Лингвиста, разрабатываемый в Совместной научно-исследовательской лаборатории «Проблемы Искусственного интеллекта» Академии наук Татарстана и Казанского университета, к блокам морфологического и синтаксического анализа [1] добавлен новый блок, называемый Синтаксическим генератором (СГ). Синтаксический генератор формирует текстовое представление татарского предложения из его синтаксической структуры. Структура, используемая в качестве исходной в синтаксическом генераторе, совпадает со структурой, получаемой на выходе синтаксического анализатора и представляет собой дерево зависимостей.

То, что из полученного синтаксическим анализатором дерева генератор порождает орфографически корректное татарское предложение, показывает, что информации, выдаваемой синтаксическим анализатором, достаточно для работы генератора.

1. Структурное описание генератора

Синтаксический генератор состоит из двух основных частей:

1. Инструмент для построения дерева зависимостей.
2. Процедуры генерации предложения из построенного дерева зависимостей.

1.1. Инструмент для создания дерева зависимостей

Инструмент для создания дерева зависимостей представляет собой диалоговое окно, позволяющее пользователю строить деревья зависимостей, узлами которого являются отдельные словоформы или аналитические конструкции (Рис.1). При этом объекты, размещенные на этом окне, исключают формирования ряда вариантов некорректных узлов дерева.

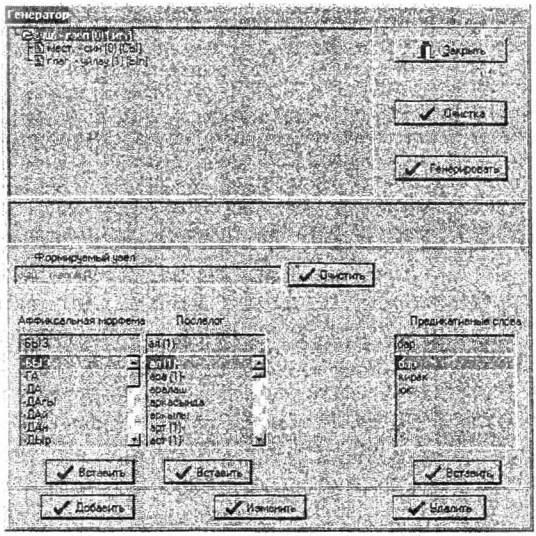

Рис.1. Окно синтаксического генератора

Диалоговое окно синтаксического генератора состоит из трех основных частей:

1. Панель, отображающая создаваемую структуру.
2. Панель, отображающая сгенерированное предложение.
3. Панель, на которой формируются аналитические конструкции или отдельные словоформы, образующие узлы дерева зависимостей.

Первая панель выглядит следующим образом (Рис.2).

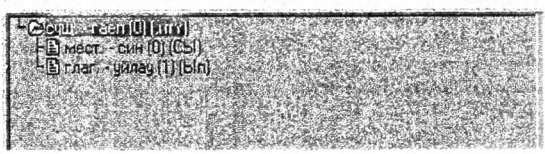

Рис.2 Панель дерева зависимостей

На панели размещен визуальный объект, представляющий собой дерево зависимостей. Узлами дерева являются некоторые структуры, описывающие как аналитические конструкции, так и отдельные словоформы.

Например:

глаг. - бару (1) (ЫП, карау, (глаг;1), ЫргА, кирәк) — структура, выражающая аналитическую конструкцию: *барып карарга кирәк 'нужно попробовать сходить'*

Здесь:

глаг. — часть речи основной словоформы в аналитической конструкции,

бару — корневая морфема основной словоформы

(1) — указание признака «твердости/ мягкости» корневой морфемы

ЫП — последовательность аффиксальных морфем основной словоформы аналитической конструкции. В данном случае это одна морфема.

карау, (глаг;1), ЫргА — соответственно, корневая морфема, указание ее части речи, признака твердости/ мягкости и последовательность аффиксальных морфем для второй словоформы в аналитической конструкции.

кирәк — третья словоформа аналитической конструкции.

сущ. - урман (1) (ГА) — структура представляющая отдельную словоформу: *урманга 'в лес'*

Например, дерево зависимостей для предложения — *Малайга урманга барып карарга кирәк. 'Мальчику нужно попробовать сходить в лес'* сформированное с помощью таких конструкций-узлов, будет иметь следующую структуру (Рис.3):

глаг. - бару (1) (ЫП, карау, (глаг;1), ЫргА, кирәк)

- **сущ. - урман (1) (ГА)**
- **сущ. - малай (1) (ГА)**

Рис.3. Дерево зависимостей

Вторая панель представляет собой текстовое окно, в котором отображается предложение (Рис.4), сгенерированное из сформированной структуры.

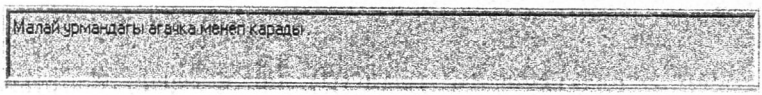

Рис.4 Панель со сгенерированным предложением

Третья панель отображает на себе инструменты, предоставляющие пользователю возможность формирования узлов дерева зависимостей (Рис.5).

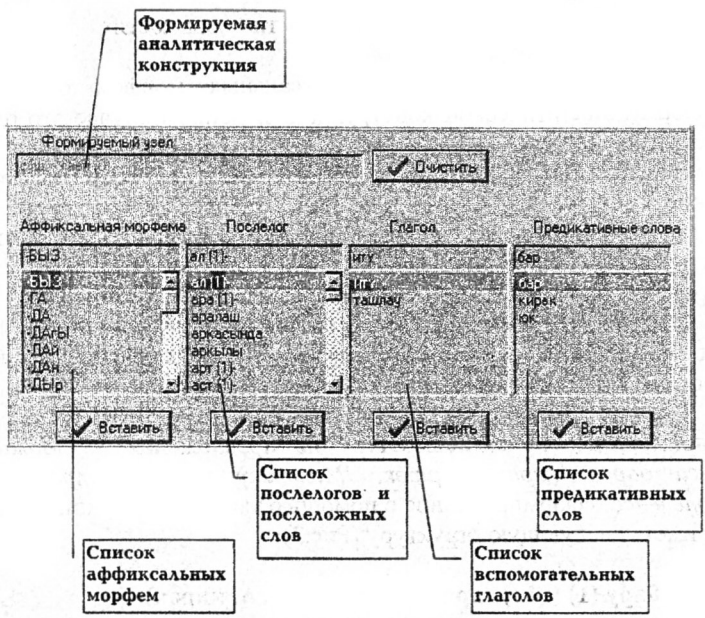

Рис.5. Панель формирования узлов дерева зависимостей

В левом верхнем углу панели расположено окно, в котором отображается формируемый узел дерева. После добавления сформированной конструкции в дерево зависимостей данное окно автоматически очищается и готово для формирования нового узла.

Формирование конструкции начинается со ввода некоторой корневой морфемы в окно, расположенное в правом верхнем углу панели. После нажатия кнопки «Вставить», расположенной рядом с окошком, производится ее поиск в словаре основ и, в случае успешного поиска, данная морфема приобретает атрибуты «Часть

речи» и «Мягкость-твердость». Затем данное окно исчезает и появляются следующие четыре списка (рис.5):

1. Список аффиксальных морфем, которые можно присоединить к последней из словоформ в полученной конструкции.
2. Послелоги и послеложные слова, которые можно добавить к полученной конструкции.
3. Вспомогательные глаголы.
4. Предикативные слова.

Список аффиксальных морфем — это список тех морфем, которые можно присоединить справа к последней морфеме, корневой или аффиксальной, в формируемом узле. От того, какой является последняя морфема в узле, корневой или аффиксальной, зависит принцип формирования списка морфем, способных присоединиться к ней справа. Корневые морфемы подразделяются на типы по частям речи и для каждого типа имеется свой список аффиксальных морфем, которые загружаются в данную колонку. Если последней морфемой в конструкции является аффиксальная морфема, то используется иной алгоритм определения списка возможных морфем. В морфологическом аспекте Модели татарской морфемы [2,3] имеется пункт «Порядок следования морфем в словоформе», где указываются все морфемы, которые могут следовать слева и справа от рассматриваемой морфемы. В данной программе используется пункт, где описываются все морфемы, следующие справа от текущей. Этот список и выводится в рассматриваемой колонке. Если для какой-либо морфемы данный список является пустым, то эта колонка на панели отсутствует. Таким образом данная колонка позволяет формировать только правильные последовательности аффиксальных морфем.

Следующие три столбца (Рис.2) содержат соответственно списки послелогов и послеложных слов, вспомогательных глаголов и предикативных слов. Алгоритм заполнения этих столбцов зависит от типа последней морфемы в узле, т.е. того какой является морфема — корневой или аффиксальной. Так, для аффиксальной морфемы элементы этих списков формируются на основе информации синтаксического аспекта Модели татарской морфемы. В этом аспекте имеется пункт «Аналитические конструкции», который для каждой словоформы, на правом конце которой находится исследуемая морфема, определяет какие элементы, формирующие аналитическую конструкцию, могут следовать в предложении справа от нее.

Пункт «Аналитические конструкции» имеет следующую структуру:

Аналитические конструкции

- **Послелоги и послеложные слова**
- **Вспомогательные глаголы**
- **Предикативные слова**

Рис.6. Структура пункта «Аналитические конструкции»

Если последней морфемой в конструкции является корневая морфема, то содержимое этих трех списков определяется тем, к какой части речи относится данная корневая морфема, поскольку для каждой части речи определены свои списки послелогов, вспомогательных глаголов и предикативных слов, с помощью которых они образуют аналитические конструкции.

Исключением является словарь вспомогательных глаголов для существительных, с помощью которых образуется аналитическая конструкция типа: **существительное + глагол.** Данное сочетание в работе [4] определяется как один составной глагол.

Примеры таких сочетаний:
җавап биру 'отвечать'
азап чигү 'мучиться'
азат иту 'освободить'

Структура этого словаря составных глаголов такова, что для каждого существительного определяется свой набор вспомогательных глаголов.

Например:

а) акыл жую 'потерять сознание'
акыл биру 'учить уму-разуму'
акыл керү 'поумнеть'
акыл сату 'поучать'

б) ачу алу 'отомстить'
ачу йоту 'проглотить обиду'
ачу кабару 'разозлиться'
ачу китерү 'злить'

1.2. Процедуры генерации предложения

После завершения построения синтаксической структуры для генерации предложения из этой структуры необходимо нажать кнопку «Генерировать». В результате чего запускаются процедуры, преобразующие предложение в поверхностное представление из его структуры.

Порядок формирования предложения основывается на следующих двух утверждениях:

1. В орфографически корректном татарском предложении предикат располагается в самом конце.
2. Каждый член предложения, зависящий от некоторого другого, располагается левее него.

Порядок нескольких актантов одного и того же предиката синтаксическими правилами татарского языка не определяется, а зависит от актуального членения предложения, т.е. от того, что в предложении является темой, а что ремой. Поскольку на данном этапе такие аспекты, как актуальность и акцентуация не рассматриваются, то порядок расположения актантов предиката совпадает с порядком расположения соответствующих им узлов в самом дереве относительно узла предиката.

Генерация предложения производится путем обхода всех узлов данного дерева, начиная с корневого узла дерева, в котором располагается предикат. В каждом узле дерева находится некоторая конструкция, которую необходимо привести к текстовому представлению. Приведение происходит разделением конструкции на представления отдельных словоформ и запуском морфологического генератора для каждой словоформы.

Так, например, из конструкции

глаг. - бару (1) (ЫП, карау, (глаг;1), ЫргА, кирәк)

после разбиения на словоформы на вход морфологического генератора подается следующая последовательность:

(бару, ЫП, глаг., 1)
(карау, ЫргА, глаг, 1)
(кирәк, , пред., 0)

В этих конструкциях первый элемент основа, второй — набор присоединяемых аффиксальных морфем, третий — часть речи основы, четвертый признак твердости или мягкости словоформы.

Морфологический генератор из этих конструкций порождает словоформы по алгоритму, представленному на рис.7:

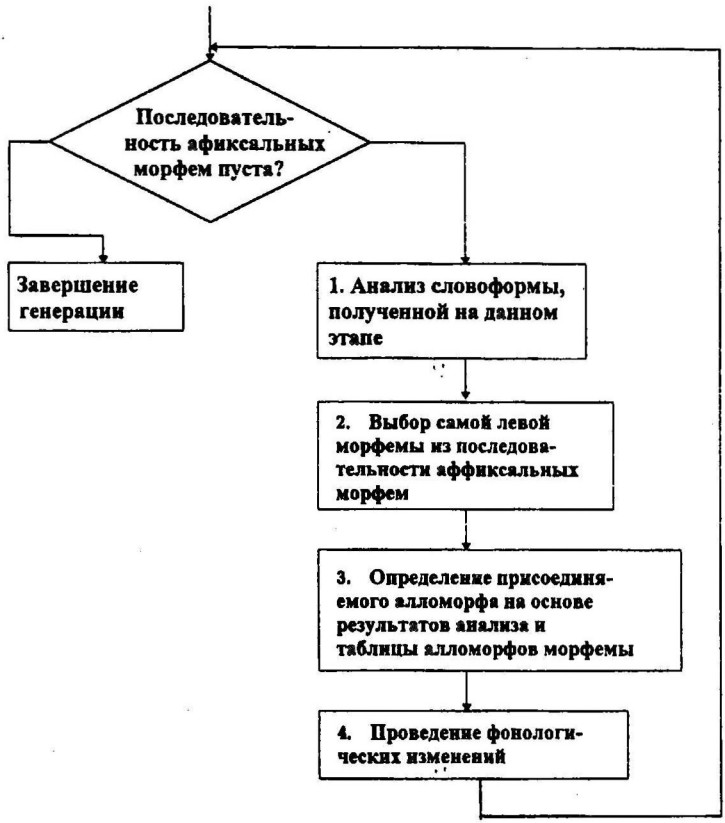

Рис.7. Алгоритм работы морфологического генератора

Рассмотрим некоторые процедуры морфологического генератора.

1. Анализ словоформы.

Анализ словоформы представляет собой процедуры определения некоторых признаков для выбора алломорфа:

- твердость/ мягкость,
- наличие притяжательности,

- тип последней буквы — гласная или согласная и какого типа.

Множества значений, которые выдает анализатор, рассмотрены в работе [5].

2. Определение алломорфа.

Определение алломорфа для соответствующей морфемы производится по таблице алломорфов для этой морфемы, путем сопоставления результатов анализа словоформы с информацией, хранящейся в таблице алломорфов. Структура и содержимое таблицы алломорфов рассмотрены в работе [5].

3. Возможные фонологические изменения.

Морфологический генератор производит фонологические изменения следующих типов:

а) Приведение глаголов из формы имени действия к форме повелительного наклонения,

б) Преобразования длЯ чередующихся букв, таких как ‘б’—‘п’ и ‘г’—‘к’,

в) Удаление некоторых символов.

Приведение глаголов к форме повелительного наклонения связано с тем, что в словаре основ глаголы хранятся в форме имени действия, а глагольные аффиксальные морфемы присоединяются к форме повелительного наклонения.

Примеры преобразований:

кайту ‘возвращение’ — кайт ‘вернись’

баю ‘обогащение’ — бае ‘богатей’

кую ‘установка’ — куй ‘ставь’

саву ‘доение’ — сау ‘дои’

Преобразования чередующихся букв можно наблюдать на следующих примерах:

китап ‘книга’ — китабым ‘моя книга’

йозак ‘замок’ — йозагым ‘мой замок’

Удаление символов производится в том случае, если для алломорфа указывается, что перед его присоединением должен быть отсечен символ.

Например:

кара ‘смотри’ — кар(а)+ый ‘смотрит’

куй ‘ставь’ — ку(й)+я ‘ставит’

Второй случай отсечения можно видеть на следующем примере: *халык 'народ' — халкым 'мой народ'*

Заключение. Синтаксический генератор, описанный в данной статье реализован в системе Delphi и может быть использован в составе системы перевода на татарский язык и системы перефразирования.

ЛИТЕРАТУРА

1. Сулейманов Д.Ш., Гатиатуллин А.Р. *Синтаксический анализатор предложений татарского языка.* // В данном сборнике.
2. Сулейманов Д.Ш., Гатиатуллин А.Р. *Модель аффиксальных морфем* // Модели национальных языков. Серия: Интеллект. Язык. Компьютер. — Казань: Изд - во 'Фән'. Вып.4. - 1996. С.113-127.
3. Сулейманов Д.Ш., Гатиатуллин А.Р. *Функционально-структурная модель татарских морфем как база данных для лингвопроцессоров.* //Сборник трудов Международного семинара по компьютерной лингвистике и ее приложениям «Диалог-97», - Ясная Поляна, 1997. - с.266-271.
4. М.З.Закиев. *Татарская грамматика.* Т.III. - Казань: Таткнигоиздат. - 1992. - 448с.
5. Гатиатуллин А.Р. *К реализации таблицы алломорфов модели татарской морфемы.* // В данном сборнике.